Supplementary – LevelSet R-CNN: A Deep Variational Method for Instance Segmentation

Namdar Homayounfar*[1,2] Yuwen Xiong*[1,2] Justin Liang*[1]
Wei-Chiu Ma[1,3] Raquel Urtasun[1,2]
{namdar,yuwen,justin.liang,weichiu,urtasun}@uber.com

[1] Uber Advanced Technologies Group
[2] University of Toronto
[3] MIT

In Section 1 of this supplementary material, we present the architectural details of our model. In section 2 we showcase more qualitative examples.

1 Architecture

The `Resnet-FPN` backbone, the bounding box head and the classification head of our model (shown in Figure 1) follow the original architecture of [1]. In our experiments, we use `Resnet-50` [2] and `WideResnet-38` [3] as the residual backbones of our model. Next, we describe the architecture details of the initial truncated signed distance function (TSDF) head, the hyperparameter head and the Chan-Vese feature head.

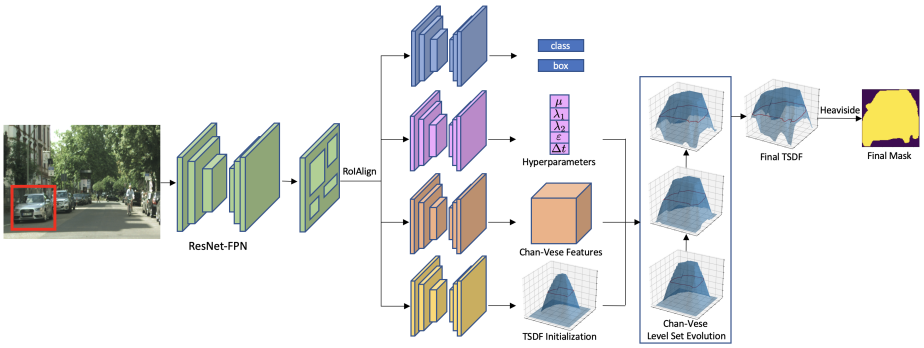

Fig. 1. Our Model for Instance Segmentation: We build on top of Mask R-CNN to first detect and classify all objects in the image. Then for each detection, the corresponding RoI is fed to a series of convolutions to obtain a truncated signed distance function (TSDF) initialization, a deep feature tensor, and a set of instance aware adaptive hyperparameters. These in turn are inputted into an unrolled Chan-Vese level set optimization procedure which outputs a final TSDF. We obtain a mask by applying the Heaviside function to the TSDF.

1.1 Initial TSDF Head

The TSDF head outputs an unrestricted TSDF for TSDF initialization as shown in Figure 1. The ground truth TSDF is truncated at a threshold of $\lfloor \frac{\text{Mask Dim}}{7} \rfloor$, i.e. 4 for a mask of 28×28. The following Table 1 shows the exact architecture.

Type	Input Size (C × H × W)	Filters @ Kernel Size	Output Size (C × H × W)
Convolution + ReLU	256 × 14 × 14	256 @ 3 × 3	256 × 14 × 14
Convolution + ReLU	256 × 14 × 14	256 @ 3 × 3	256 × 14 × 14
Convolution + ReLU	256 × 14 × 14	256 @ 3 × 3	256 × 14 × 14
Convolution + ReLU	256 × 14 × 14	256 @ 3 × 3	256 × 14 × 14
DeConvolution + ReLU	256 × 14 × 14	256 @ 2 × 2	256 × 28 × 28
Convolution	256 × 28 × 28	8 @ 1 × 1	8 × 28 × 28

Table 1. The initial TSDF head architecture.

1.2 Hyperparameter Head

The hyperparameter head outputs a vector of size $2N + 3$ for hyperparameters $\{\mu(r_m), \lambda_1(r_m), \lambda_2(r_m), \varepsilon_{1:N}(r_m), \Delta t_{1:N}(r_m)\}$ for each RoI r_m as shown in Figure 1. Here N corresponds to the number of unrolled optimization steps. In Table 2 we present the exact architecture.

Type	Input Size (C × H × W)	Filters @ Kernel Size	Output Size (C × H × W)
Convolution + ReLU	256 × 28 × 28	256 @ 3 × 3	256 × 14 × 14
Avg Pooling	256 × 14 × 14		256 × 1 × 1
Convolution + ReLU	256 × 1 × 1	256 @ 1 × 1	256 × 1 × 1
Convolution + ReLU	256 × 1 × 1	128 @ 1 × 1	128 × 1 × 1
Convolution + 2 × Sigmoid	128 × 1 × 1	$2N+3$ @ 1 × 1	$(2N+3)$ × 1 × 1

Table 2. The hyperparameter head architecture.

1.3 Chan-Vese Features Head

In Table 3 we showcase the exact architecture of the Chan-Vese feature head. In our experiments we found the channel size of 64 and dimensions of 112 to be effective in terms of memory.

Type	Input Size (C × H × W)	Filters @ Kernel Size	Output Size (C × H × W)
Convolution + ReLU	256 × 56 × 56	64 @ 3 × 3	64 × 56 × 56
Convolution + ReLU	64 × 56 × 56	64 @ 3 × 3	64 × 56 × 56
Convolution + ReLU	64 × 56 × 56	64 @ 3 × 3	64 × 56 × 56
DeConvolution + ReLU	64 × 56 × 56	64 @ 2 × 2	64 × 112 × 112
Convolution	64 × 112 × 112	64 @ 3 × 3	64 × 112 × 112

Table 3. The Chan-Vese features head architecture.

2 Qualitative Results

In Figures 2 and 3 we showcase the input image, our state-of-the art results with the `WideResnet-38` backbone (without test time augmentation or COCO pre-training) and the corresponding GT on the Cityscapes validation set. In Figures 4 and 5 we showcase our model with the `Resnet-50` backbone on the validation set of COCO dataset.

Input Image | Our Instance Segmentation | GT Instance Segmentation

Fig. 2. We showcase qualitative instance segmentation results of our model on the Cityscapes validation set.

Fig. 3. We showcase qualitative instance segmentation results of our model on the Cityscapes validation set.

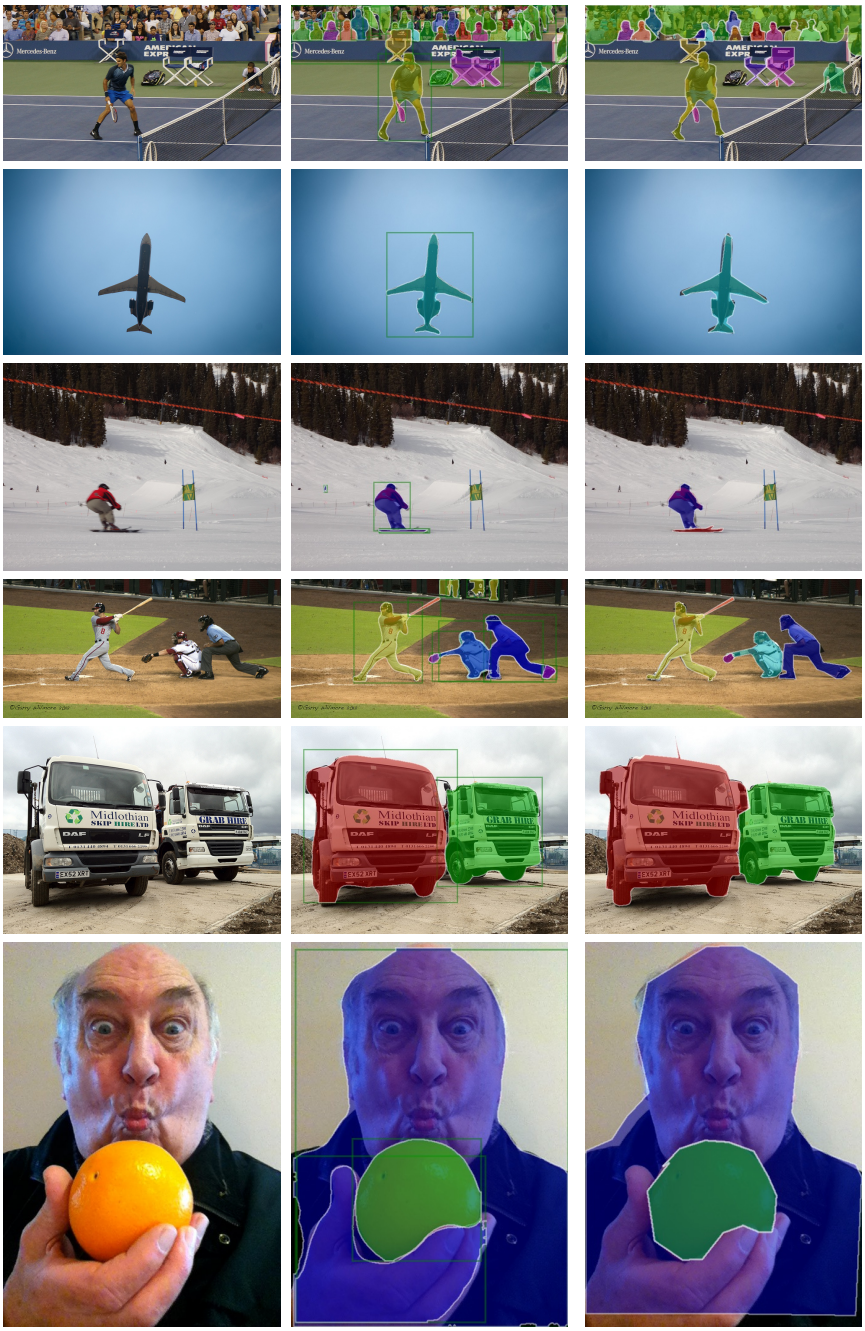

Fig. 4. We showcase qualitative instance segmentation results of our model on the COCO validation set.

Fig. 5. We showcase qualitative instance segmentation results of our model on the COCO validation set.

References

1. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: CVPR (2017)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CVPR (2015)
3. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: BMVC (2016)